

NGHIÊN CỨU VÀ XÂY DỰNG CẤU TRÚC TÔPÔ THEO TIẾP CẬN TẬP THÔ, ỨNG DỤNG RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH

Trần Thanh Đại¹, Nguyễn Long Giang², Vũ Đức Thi³, Trần Thị Huệ⁴

¹Trường Đại học Kinh tế Kỹ thuật Công nghiệp

²Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

³Viện Công nghệ thông tin, Đại học Quốc gia Hà Nội

⁴Trường Đại học Điện lực

tt daiuneti@gmail.com, nlgang@ioit.ac.vn, vdthi@vnu.edu.vn, dinhthukhanh@gmail.com, huongtrieu@hvn.edu.vn, huett@dainam.edu.vn, tvha@ictu.edu.vn, dungkt@tlu.edu.vn, longck.2006@gmail.com

TÓM TẮT: Rút gọn thuộc tính là bước tiền xử lý dữ liệu quan trọng trong các lĩnh vực học máy và khai thác dữ liệu. Trong những năm gần đây, mối quan hệ giữa mô hình lý thuyết tập thô và tôpô cũng như các phương pháp xây dựng cấu trúc tôpô theo tiếp cận tập thô nhận được nhiều quan tâm của các nhà nghiên cứu. Tuy nhiên, phương pháp xây dựng cấu trúc tôpô theo tiếp cận tập thô cho bài toán rút gọn thuộc tính còn gặp nhiều thách thức. Trong nghiên cứu này, chúng tôi đề xuất xây dựng cấu trúc tôpô dựa trên mối quan hệ thứ tự của các phép toán xấp xỉ trên và xấp xỉ dưới của tập thô mờ. Trên cơ sở đó, chúng tôi đề xuất phương pháp rút gọn thuộc tính theo tiếp cận tính toán biên của tôpô đề xuất, ứng dụng cho bài toán rút gọn thuộc tính.

Từ khóa: Tôpô, Tập thô, Rút gọn thuộc tính, Bảng quyết định.

I. GIỚI THIỆU

Rút gọn thuộc tính là bước tiền xử lý dữ liệu quan trọng, ứng dụng rộng rãi trong các lĩnh vực học máy và khai thác dữ liệu. Mục tiêu chung của bài toán rút gọn thuộc tính là đi tìm một tập con của tập thuộc tính ban đầu thỏa mãn một ràng buộc cho trước. Đối với bảng quyết định, tập con phải thỏa mãn tiêu chí bảo toàn thông tin quyết định so với tập thuộc tính gốc. Bài toán đi tìm tập con của tập thuộc tính ban đầu là một bài toán NP-Hard, do đó các phương pháp rút gọn thuộc tính hiện nay vẫn còn gặp nhiều thách thức.

Mô hình lý thuyết tập thô (Rough Set - RS) do Pawlack đề xuất vào năm 1982 đã được chứng minh là một trong những công cụ toán học mạnh mẽ, ứng dụng hiệu quả cho các bảng dữ liệu không nhất quán, không chắc chắn và không đầy đủ. Cho đến nay đã có nhiều phương pháp rút gọn thuộc tính theo tiếp cận tập thô và tập thô suy rộng được các nhà nghiên cứu phát triển và đề xuất [1-5].

Khái niệm không gian tôpô trên nền tập thô lần đầu tiên được đưa ra bởi Pawlack và cộng sự. Trong đó, các phân hoạch được xem như là một cơ sở (base) của tôpô. Khi đó, Pawlack và các cộng sự cũng đã khẳng định cấu trúc đại số của tập thô với tôpô là tương đương trên nền tập thô với các phép toán xấp xỉ dưới của tập thô tương đương với toán tử miền trong (INT) của tôpô và phép toán xấp xỉ trên của tập thô tương đương với toán tử miền ngoài (CLS) của tôpô.

Trên cơ sở các đề xuất ban đầu của Pawlack và các cộng sự, Lashin và các cộng sự đưa ra khái niệm biểu diễn tri thức trên không gian tôpô và phương pháp loại bỏ các tri thức dư thừa thông qua bảo toàn cấu trúc tôpô ban đầu [6]. Tiếp đó Su đề xuất khái niệm không gian tôpô trên họ các phủ của tập thô để rút gọn thuộc tính trong bảng quyết định không đầy đủ [13]. Tuy nhiên các nghiên cứu vẫn dừng lại ở giai đoạn đề xuất mà chưa có ứng dụng trên các bộ dữ liệu thực và đánh giá tính hiệu quả của mô hình đề xuất. Các phương pháp chính để xây dựng cấu trúc tôpô theo tiếp cận tập thô gồm có:

Phương pháp sinh tôpô dựa trên không gian xấp xỉ: Đối với mô hình tập thô truyền thống, không gian xấp xỉ là các phân hoạch hay tập các lớp tương đương. Do đó, các phân hoạch này luôn là các cơ sở (base) của một tôpô. Khi đó, để nghiên cứu cấu trúc tôpô, các nhà nghiên cứu thường phân tích cấu trúc của một base là đủ thông tin cần phân tích [6, 14]. Do đó, khi một quan hệ xây dựng không gian xấp xỉ có tính chất khác nhau thì base cũng sẽ khác nhau. Đối với các phủ sinh bởi các quan hệ dung sai thì các phủ này được coi là các subbase, khi đó dựa trên các tính chất của base sinh từ subbase, có thể đề xuất các phương pháp xây dựng base từ các phủ này [13]. Trên nền tập mờ (Fuzzy Set - FS) các nhà nghiên cứu cũng đề xuất phương pháp xây dựng tôpô từ các không gian xấp xỉ mờ [10].

Phương pháp sinh tôpô dựa trên các phép toán xấp xỉ: Mô hình tập thô có hai phép toán cơ bản là xấp xỉ trên và xấp xỉ dưới để đánh giá mức độ thô của một tập trong một không gian xấp xỉ. Pawlack và cộng sự cũng đã chỉ ra mối quan hệ giữa tập xấp xỉ trên và dưới có tính chất thứ tự, trong đó tập xấp xỉ trên luôn lớn hơn tập xấp xỉ dưới. Do đó, kết hợp các tập xấp xỉ dưới và tập xấp xỉ trên luôn thỏa mãn là một cấu trúc tôpô [12, 14, 15].

Phương pháp sinh tôpô dựa trên không gian mẫu: Dựa trên mối quan hệ giữa các phép toán và không gian mẫu (đối tượng) trong RS, các nhà nghiên cứu đề xuất các phương pháp khác nhau để xây dựng cấu trúc tôpô. Dựa trên mối quan hệ này, các tác giả đã chứng minh họ các tập con của không gian mẫu là một cấu trúc tôpô nếu các tập con này là các tập xác định (tập chính xác) trong RS [7, 8, 11].

Gần đây có công trình nghiên cứu của Xie và các cộng sự có đề xuất hướng rút gọn thuộc tính dựa trên ma trận phân biệt sử dụng cấu trúc tôpô, trong đó tính chất phân biệt được mở rộng dựa trên cấu trúc tôpô [9]. Tuy nhiên, các phương pháp xây dựng cấu trúc tôpô các tác giả đề xuất vẫn còn nhiều hạn chế về thời gian xây dựng. Trong nghiên cứu này, chúng tôi đề xuất cấu trúc tôpô dựa trên quan hệ thứ tự của các phép toán xấp xỉ trên và xấp xỉ dưới. Trên cơ sở đó, chúng tôi đề xuất phương pháp rút gọn thuộc tính theo tiếp cận tính toán biên của tôpô đề xuất.

Ngoài phần giới thiệu, phần còn lại của bài báo có cấu trúc như sau. Mục II nhắc lại một số kiến thức cơ bản, Mục III đề xuất cấu trúc tôpô theo tiếp cận tập thô trên không gian xấp xỉ mờ. Mục IV đề xuất mô hình rút gọn thuộc tính cho bảng quyết định theo tiếp cận tôpô và một số kết quả thực nghiệm được trình bày trong Mục V. Cuối cùng là phần kết luận và hướng phát triển của bài báo.

II. KIẾN THỨC NỀN TẢNG

Phần này nhắc lại một số kiến thức cơ bản về bảng quyết định, tập thô, tập mờ và không gian tôpô.

A. Tập thô và tập mờ

1. Quan hệ tương đương và các phân hoạch trên bảng quyết định

Bảng quyết định là một hệ thông tin với tập thuộc tính gồm có hai tập thuộc tính khác rỗng rời nhau gồm C và D , lần lượt được gọi là tập thuộc tính điều kiện và tập thuộc tính quyết định kí hiệu bởi $DT = (U, C, D)$ với $C \cap D = \emptyset$.

Xét bảng quyết định $DT = (U, C, D)$. Mỗi tập con các thuộc tính $P \subseteq C$ xác định một quan hệ hai ngôi trên U , ký hiệu là $R(P)$, xác định bởi

$$R(P) = \{(x, y) \in U \times U \mid \forall a \in P, a(x) = a(y)\}.$$

Khi đó $R(P)$ được gọi là quan hệ không phân được theo P hay (P -không phân biệt được). Rõ ràng $R(P)$ là một quan hệ tương đương trên U . Do đó, nếu $(x, y) \in R(P)$ thì hai đối tượng x và y không phân biệt được bởi các thuộc tính trong P . Quan hệ tương đương $R(P)$ xác định một phân hoạch trên U , ký hiệu là $U / R(P)$ hay U / P . Ký hiệu lớp tương đương trong phân hoạch U / P chứa đối tượng x là $[x]_P$, khi đó $[x]_P = \{y \in U \mid (x, y) \in R(P)\}$.

2. Mô hình tập thô

Trong lý thuyết tập thô, để đánh giá các đối tượng trong tập $X \subseteq U$, người ta lấy xấp xỉ X với U / P thông qua các phương pháp xấp xỉ. Có hai phương pháp lấy xấp xỉ của X với U / P gồm có *xấp xỉ dưới* và *xấp xỉ trên*, ký hiệu lần lượt là $\underline{P}(X)$ và $\overline{P}(X)$, được xác định như sau:

$$\underline{P}(X) = \{x \in U \mid [x]_P \subseteq X\}, \quad \overline{P}(X) = \{x \in U \mid [x]_P \cap X \neq \emptyset\}.$$

3. Tập mờ

Cho U là tập hữu hạn khác rỗng các đối tượng, tập mờ A trên U có dạng: $A = \{\langle x, \mu_A \rangle \mid x \in U\}$ với $\mu_A : U \rightarrow [0, 1]$ thỏa mãn $0 \leq \mu_A(x) \leq 1$. Trong đó $\mu_A(x)$ được gọi là độ thành viên của phần tử x trong A .

B. Không gian tôpô

1. Các khái niệm cơ bản

Định nghĩa 2.1. [Không gian tôpô]

Không gian tôpô [6] τ trên U là họ các tập con của U sao cho τ thỏa mãn các điều kiện sau đây

- 1) $\emptyset, U \in \tau$
- 2) $\forall G_1, G_2 \in \tau$, tồn tại $G_3 \in \tau : G_3 = G_1 \cap G_2$
- 3) $\forall G_1, G_2 \in \tau$, tồn tại $G_3 \in \tau : G_3 = G_1 \cup G_2$

Khi đó, cặp (U, τ) được gọi là không gian tôpô với τ xác định trên U . Mỗi phần tử trong τ được gọi là một tập mở (openset), bù của tập mở được gọi là tập đóng (closedset).

Định nghĩa 2.2. [Cơ sở của không gian tôpô]

Cơ sở (base) của tôpô [6] τ trên U là họ các tập con B của U sao cho với mọi $G_1, G_2 \in B$ thì $G_1 \cap G_2 \in B$.

2. Cấu trúc tôpô theo tiếp cận tập thô

Định nghĩa 2.3. [Cấu trúc tôpô thô]

Cho $\underline{P}(X)$ và $\overline{P}(X)$ là các tập xấp xỉ trên và xấp xỉ dưới của X trên phân hoạch P . Khi đó: $\tau = \{\underline{P}(X), \overline{P}(X)\}$ được gọi là tôpô thô của tập X [8].

III. CẤU TRÚC TÔPÔ TRÊN KHÔNG GIAN XẤP XỈ MỜ

Phần này, chúng tôi đề xuất cấu trúc tôpô mờ dựa trên hai thành phần chính của tập thô mờ đó là tập xấp xỉ mờ trên và tập xấp xỉ mờ dưới của một tập X cho trước.

A. Không gian xấp xỉ mờ

Định nghĩa 3.1. [Quan hệ mờ]

Cho U là tập khác rỗng các đối tượng, quan hệ quan hệ mờ $\tilde{R}$ của hai đối tượng $x, y \in U$ được xác định bởi:

$$\tilde{R}(x, y) = 1 - |V(x) - V(y)| \quad (1)$$

trong đó: $V(x), V(y)$ tương ứng là giá trị biểu diễn của đối tượng $x, y \in U$ với $V(x), V(y) \in [0, 1]$.

Định nghĩa 3.2. [Không gian xấp xỉ mờ]

Cho U là tập khác rỗng các đối tượng và quan hệ mờ $\tilde{R}$ xác định trên U . Khi đó, cặp $(U, \tilde{R})$ được gọi là không gian xấp xỉ mờ.

Định nghĩa 3.3. [So sánh các quan hệ]

Cho $\tilde{P}, \tilde{Q} \in (U, \tilde{R})$. Khi đó, $\tilde{P}$ được gọi là mịn hơn $\tilde{Q}$ nếu $\tilde{P}(x, y) \leq \tilde{Q}(x, y)$ với mọi $x, y \in U$.

Định nghĩa 3.4. [Quan hệ mịn nhất]

Cho $\tilde{P} \in (U, \tilde{R})$. Khi đó, $\tilde{P}$ được gọi là mịn nhất nếu $\tilde{P}(x, x) = 1$ và $\tilde{P}(x, y) = 0$ với mọi $x, y \in U$.

B. Tôpô trên không gian xấp xỉ mờ

Định nghĩa 3.5. [Tôpô mờ]

Cho $\tilde{P} \in (U, \tilde{R})$ và $X \subseteq U$. Khi đó $T_{\tilde{P}} = \{\underline{X}_{\tilde{P}}, \overline{X}_{\tilde{P}}\}$ là một tôpô mờ trên X . Trong đó:

$$\begin{aligned} \underline{X}_{\tilde{P}} &= \left\{ (x, \underline{X}(x)) : \underline{X}(x) = \bigwedge_{y \in U} I(\tilde{P}(x, y), X(x)), \forall x \in U \right\} \\ \overline{X}_{\tilde{P}} &= \left\{ (x, \overline{X}(x)) : \overline{X}(x) = \bigvee_{y \in U} T(\tilde{P}(x, y), X(x)), \forall x \in U \right\} \end{aligned}$$

Nhận thấy rằng các tập $\underline{X}_{\tilde{P}}$ và $\overline{X}_{\tilde{P}}$ là các tập mờ với U phần tử được mờ hóa trên X . Trong đó các toán tử kéo theo I và toán tử T-norm T trong logic mờ được sử dụng.

Định nghĩa 3.6. [So sánh tôpô]

Cho $T_{\tilde{P}} = \{\underline{X}_{\tilde{P}}, \overline{X}_{\tilde{P}}\}$ và $T_{\tilde{Q}} = \{\underline{X}_{\tilde{Q}}, \overline{X}_{\tilde{Q}}\}$ xác định trên X là hai tôpô mờ tương ứng với $\tilde{P}, \tilde{Q} \in (U, \tilde{R})$. Khi đó $T_{\tilde{P}}$ được gọi là rõ hơn $T_{\tilde{Q}}$ nếu $d_E(\underline{X}_{\tilde{P}}, \overline{X}_{\tilde{P}}) < d_E(\underline{X}_{\tilde{Q}}, \overline{X}_{\tilde{Q}})$. Trong đó, d_E là khoảng cách Euclid.

Mệnh đề 3.1. [So sánh tôpô qua quan hệ]

Cho $T_{\tilde{P}} = \{\underline{X}_{\tilde{P}}, \overline{X}_{\tilde{P}}\}$ và $T_{\tilde{Q}} = \{\underline{X}_{\tilde{Q}}, \overline{X}_{\tilde{Q}}\}$ xác định trên X là hai tôpô mờ tương ứng với $\tilde{P}, \tilde{Q} \in (U, \tilde{R})$. Khi đó $T_{\tilde{P}}$ được gọi là rõ hơn $T_{\tilde{Q}}$ nếu $\tilde{P}$ mịn hơn $\tilde{Q}$.

Định nghĩa 3.7. [Tôpô rõ nhất]

Cho $T_{\tilde{P}} = \{\underline{X}_{\tilde{P}}, \bar{X}_{\tilde{P}}\}$ xác định trên X là một tôpô mờ tương ứng với $\tilde{P} \in (U, \tilde{R})$ và khoảng cách Euclid d_E . Khi đó, $T_{\tilde{P}}$ được gọi là rõ nhất $d_E(\underline{X}_{\tilde{P}}, \bar{X}_{\tilde{P}}) = 0$. Trong đó d_E là khoảng cách Euclid

Mệnh đề 3.2. [Tôpô rõ nhất qua quan hệ mịn nhất]

Cho $T_{\tilde{P}} = \{\underline{X}_{\tilde{P}}, \bar{X}_{\tilde{P}}\}$ là một tôpô mờ tương ứng với $\tilde{P} \in (U, \tilde{R})$ trên X . Khi đó với mọi $X \subseteq U$, $T_{\tilde{P}}$ được gọi là rõ nhất nếu $\tilde{P}$ mịn nhất.

IV. RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH THEO TIẾP CẬN TÔPÔ THỎ MỜ

Hầu hết các thuật toán rút gọn thuộc tính hiện nay đều xây dựng độ đo dựa trên tập xấp xỉ dưới và không gian xấp xỉ mờ để đo lường sự phụ thuộc của thuộc tính. Tuy nhiên, trong các trường hợp dữ liệu phức tạp, lân cận của một đối tượng thường xuất hiện cả miền trong và miền ngoài của tập mục tiêu. Do đó, phần này chúng tôi đề xuất phương pháp đánh giá độ quan trọng của thuộc tính theo tiếp cận tính toán biên.

A. Mô hình rút gọn thuộc tính

Định nghĩa 4.1. [Thuộc tính quan trọng]

Cho bảng quyết định $DT = (U, C, D)$ và $B \subseteq C$. Khi đó, độ quan trọng của thuộc tính $c \in C - B$ với B được xác định như sau:

$$Sig_B(c) = \frac{1}{|D|} \sum_{X \in D} d_E(\underline{X}_{\tilde{P}}, \bar{X}_{\tilde{P}}) \quad (2)$$

với $\tilde{P} = \tilde{B} \cap \tilde{c}$, trong đó $\tilde{B}, \tilde{c}$ là các phân hoạch mờ tương ứng của tập thuộc tính B và thuộc tính c .

Định nghĩa 4.2. [Tập rút gọn]

Cho bảng quyết định $DT = (U, C, D)$ và $B \subseteq C$. Khi đó, B được gọi là tập rút gọn của C nếu:

- 1) $d_E(\underline{D}_{\tilde{B}}, \bar{D}_{\tilde{B}}) = d_E(\underline{D}_{\tilde{C}}, \bar{D}_{\tilde{C}})$;
- 2) $\forall A \subset B; d_E(\underline{D}_{\tilde{A}}, \bar{D}_{\tilde{A}}) = d_E(\underline{D}_{\tilde{C}}, \bar{D}_{\tilde{C}})$.

Trong đó d_E là khoảng cách Euclid.

B. Thuật toán đề xuất

Dựa trên các Định nghĩa 4.1 và 4.2, phần này chúng tôi đề xuất thuật toán rút gọn thuộc tính cho bảng quyết định đầy đủ bao gồm các bước như sau:

Thuật toán A_FT : rút gọn thuộc tính trong bảng quyết định số đầy đủ theo tiếp cận tôpô mờ

Input Bảng quyết định $DT = (U, C, D)$

Output: Tập rút gọn B

Giai đoạn khởi tạo

Begin

1. $B = \{\}$;

2. $\tilde{B} = [1]$; với $[1]$ là ma trận vuông hệ số bằng 1 với kích thước bằng $|U| \times |U|$

3. Tính $\tilde{C}$ theo công thức 1;

Giai đoạn tìm tập rút gọn

4. While $d_E(\underline{D}_{\tilde{B}}, \bar{D}_{\tilde{B}}) \neq d_E(\underline{D}_{\tilde{C}}, \bar{D}_{\tilde{C}})$

5. $t = \{\}$;

6. For all $c \in C - B$

7. If $Sig_B(c) \succ Sig_B(t)$ then

8. $t = c$;

9. EndIf

10. EndFor

11. $B = B \cup t$;

```
12          $\tilde{B} = \tilde{B} \cap t$  ;
13     EndWhile
14     Return  $B$  ;
End.
```

Tiếp theo chúng tôi tiến hành đánh giá độ phức tạp của thuật toán. Trước tiên chúng tôi kí hiệu $|U|, |C|, |D|$ lần lượt là số các đối tượng, số thuộc tính điều kiện và số phân lớp trong bảng quyết định $DT = (U, C, D)$. Độ phức tạp của thuật toán A_FT tại gian đoạn khởi tạo là $O(|U|^2 |C|)$. Độ phức tạp của thuật toán A_FT tại giai đoạn tìm tập rút gọn là $O(|U|^2 |C|^2)$. Khi đó, độ phức tạp của thuật toán A_FT trong trường hợp xấu nhất là $O(|U|^2 |C|^2)$.

V. MỘT SỐ KẾT QUẢ THỰC NGHIỆM

A. Kế hoạch và môi trường thực nghiệm

Mục tiêu của thực nghiệm nhằm đánh giá tính hiệu quả của thuật toán đề xuất bằng cách so sánh thời gian thực hiện, khả năng phân loại chính xác và kích thước tập rút gọn thu được từ thuật toán đề xuất A_FT với các thuật toán F_FPOS [1], F_FE [2], và F_FD [3]. Trong đó, thuật toán F_FPOS sử dụng độ đo miền dương mờ, thuật toán F_FE sử dụng độ đo Entropy mờ, thuật toán F_FD sử dụng độ đo khoảng cách mờ. Tất cả các thuật toán trên đều được xây dựng theo tiếp cận Filter thuộc tính và sử dụng tập nền FS để xây dựng các hạt thông tin mờ.

Môi trường thực nghiệm cho các thuật toán được thực hiện trên nền tảng hệ điều hành Window 10, bộ xử lý CPU core i5 Processor với bộ nhớ RAM 8Gb. Tất cả các thuật toán đều được cài đặt với ngôn ngữ lập trình Python và thực nghiệm trên 08 bộ dữ liệu thực từ UCI [12]. Chi tiết thông tin mô tả về các tập dữ liệu được trình bày trong Bảng 1, trong đó $|U|$ cho biết số lượng đối tượng, $|C|$ cho biết số thuộc tính điều kiện và $|D|$ cho biết số phân lớp của thuộc tính quyết định trong mỗi tập dữ liệu.

Bảng 1. Mô tả các bộ dữ liệu thực nghiệm

ID	Data	Describe	$ U $	$ C $	$ D $
1	Wine	Wine	178	13	3
2	Heart	Statlog (Heart)	270	13	2
3	BCWD	Breast Cancer Wisconsin (Diagnostic)	569	30	2
4	BCWP	Breast Cancer Wisconsin (Prognostic)	198	33	2
5	IS	Ionosphere	351	34	2
6	UFDC	Ultrasonic flowmeter diagnostics (C)	181	43	4
7	Sonar	Connectionist Bench	208	60	2
8	VRB	Voice Rehabilitation(Binary)	126	310	2

Các tiêu chí đánh giá cho các thuật toán gồm có: 1) Kích thước của tập rút gọn; 2) Khả năng phân loại chính xác (*accuracy* %); 3) Thời gian thực hiện của thuật toán (s). Trong đó khả năng phân loại chính xác được đo lường trên hai mô hình phân lớp SVM và KNN($K=|D|$, $p=2$), độ đo độ chính xác *accuracy* kết hợp với phương pháp đánh giá chéo 10-fold được sử dụng trên từng mô hình. Các tập dữ liệu trước khi rút gọn được chuẩn hóa về miền giá trị [0, 1].

B. Đánh giá thuật toán A_FT

Bảng 2 so sánh kích thước của tập rút gọn giữa thuật toán đề xuất A_FT so với nhóm các thuật toán khác. Như chúng ta có thể quan sát thấy tại Bảng 2, thuật toán đề xuất A_FT cho tập rút gọn có kích thước trung bình tốt hơn trên hầu hết các bộ dữ liệu so với các thuật toán khác. Mặc dù kích thước trung bình của các tập rút gọn thu được từ thuật toán đề xuất không chênh lệch đáng kể so với thuật toán [1] trên các bộ dữ liệu có số thuộc tính lớn như Sona và VRB, nhưng tốt hơn đáng kể so với các thuật toán [2] và [3].

Bảng 2. So sánh kích thước giữa các tập rút gọn

ID	Data	A_FT	F_POS [1]	F_FD [3]	F_FE [2]	$ C $
1	Wine	6	11	13	11	13
2	Heart	9	13	13	13	13
3	BCWD	10	16	21	7	30
4	BCWP	5	14	22	18	33
5	IS	18	23	29	17	34
6	UFDC	8	6	16	5	43
7	Sonar	8	10	35	12	60
8	VRB	10	12	43	22	310

Thuật toán F_POS [1] sử dụng độ đo miền dương mờ dựa trên các tập xấp xỉ dưới, trong khi đó thuật toán A_FT đo lường sự khác biệt giữa tập xấp xỉ dưới và tập xấp xỉ trên. Cả hai thuật toán này đều có đặc điểm chung là đều sử dụng tập xấp xỉ dưới. Do đó, kích thước tập rút gọn của hai thuật toán này khá giống nhau là có thể lý giải được. Tuy nhiên thuật toán A_FT kết hợp cả tập xấp xỉ trên và tập xấp xỉ dưới để đánh giá mức độ quan trọng của thuộc tính, do đó kích thước tập rút gọn của thuật toán đề xuất đặc biệt tốt trên một số bộ dữ liệu như Wine và BCWP.

Bảng 3 và Bảng 4 so sánh khả năng phân loại chính xác của các mô hình SVM và KNN trên các tập rút gọn thu được từ các thuật toán. Chúng ta có thể thấy thuật toán đề xuất cho tập rút gọn hiệu quả hơn so với các thuật toán khác về khả năng phân loại chính xác trên cả hai mô hình phân lớp. Đặc biệt cải thiện khả năng phân loại chính xác trên bộ dữ liệu có độ chính xác ban đầu thấp như UFDC.

Bảng 3. So sánh khả năng phân loại chính xác giữa các tập rút gọn trên mô hình SVM

ID	Data	A_FT	F_POS [1]	F_FD [3]	F_FE [2]	C
1	Wine	97	97	98	93	98
2	Heart	84	84	84	84	84
3	BCWD	94	96	97	92	98
4	BCWP	76	76	77	76	78
5	IS	88	84	88	86	88
6	UFDC	56	35	42	43	43
7	Sonar	65	64	67	72	65
8	VRB	84	67	86	68	83

Như đã trình bày trong Mục III và Mục IV, trong nghiên cứu này chúng tôi sử dụng mức độ chênh lệch (biên) của hai tập xấp xỉ trên và dưới để đánh giá độ quan trọng của thuộc tính. Khi biên của thuộc tính đó càng nhỏ, thuộc tính đó càng quan trọng và ngược lại. Phương pháp này cho phép chúng ta đánh giá cả miền trong lẫn miền ngoài của một thuộc tính điều kiện với thuộc tính quyết định. Do đó, phương pháp đánh giá này cho phép lựa chọn thuộc tính chính xác hơn cũng là điều có thể lý giải.

Bảng 4. So sánh khả năng phân loại chính xác giữa các tập rút gọn trên mô hình KNN.

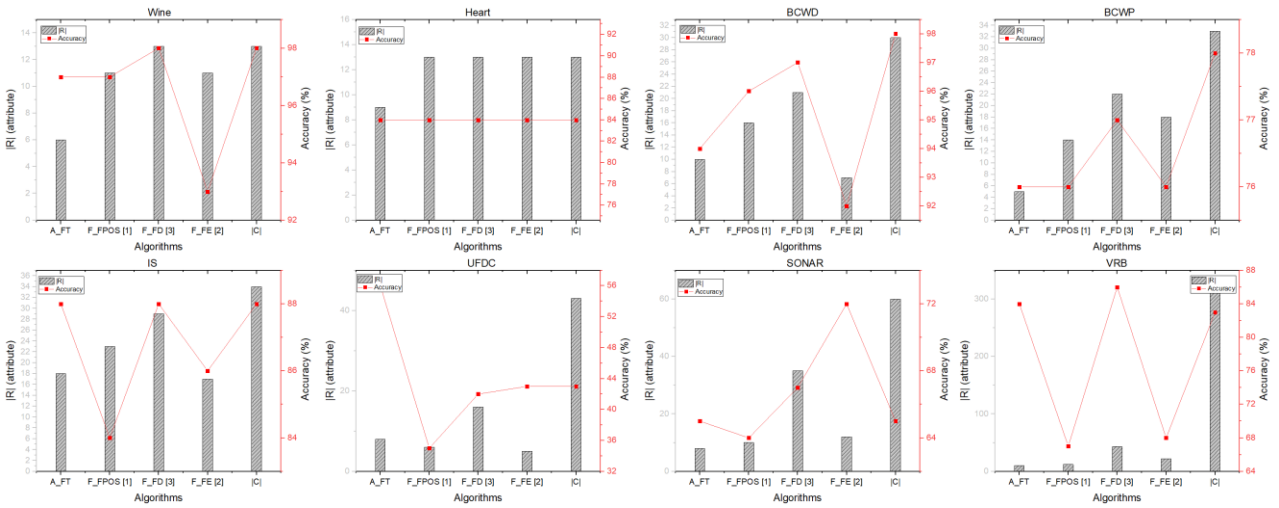
ID	Data	A_FT	F_POS [1]	F_FD [3]	F_FE [2]	C
1	Wine	98	96	98	88	96
2	Heart	80	77	84	77	77
3	BCWD	95	93	97	90	95
4	BCWP	80	79	77	78	77
5	IS	86	89	88	88	86
6	UFDC	83	68	55	70	79
7	Sonar	60	60	67	61	68
8	VRB	71	56	86	64	67

Tuy nhiên phương pháp đánh giá này cũng phải trả giá về thời gian tính toán là cao hơn nhiều so với các thuật toán khác. Quan sát Bảng 5, chúng ta có thể thấy thuật toán đề xuất A_FT cao hơn gấp đôi so với các thuật toán khác.

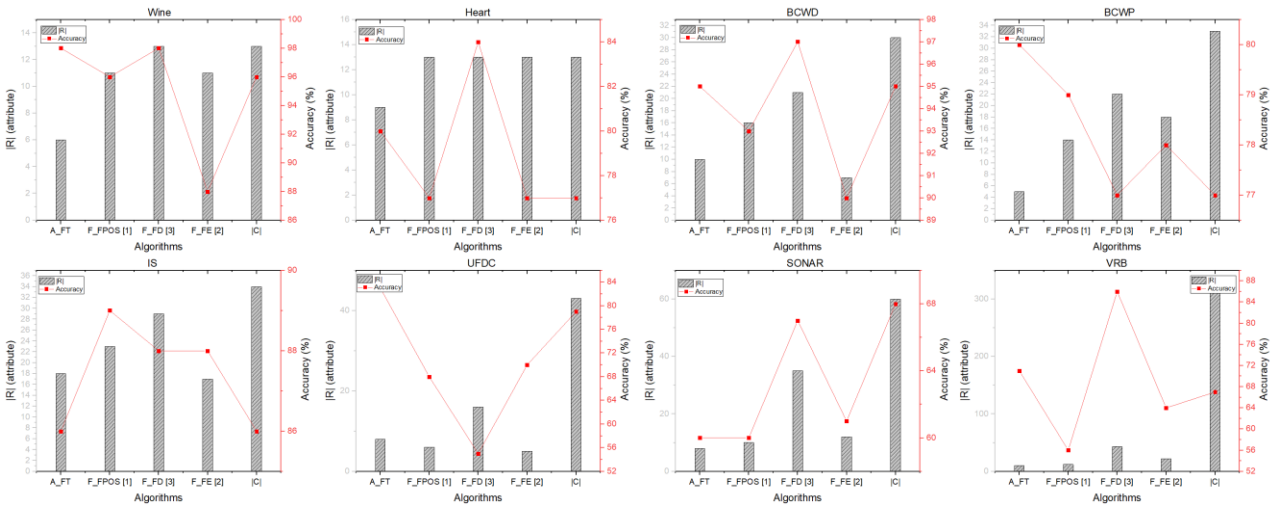
Bảng 5. So sánh thời gian tính toán tập rút gọn giữa các thuật toán.

ID	Data	A_FT	F_POS [1]	F_FD [3]	F_FE [2]
1	Wine	1.3	0.6	1.63	0.48
2	Heart	2.3	0.27	1.49	1.15
3	BCWD	5.8	1.9	5.27	3.84
4	BCWP	8.5	2.8	6.86	4.21
5	IS	5.57	0.37	5.42	3.37
6	UFDC	4.7	0.28	3.85	2.9
7	Sonar	5.7	0.24	4.63	3.31
8	VRB	1.3	0.6	1.63	0.48

Hình 1 và Hình 2 mô tả mối quan hệ về kích thước của tập rút gọn thu được từ các thuật toán với khả năng phân loại chính xác trên hai mô hình phân lớp SVM và KNN. Chúng ta có thể thấy, trên một số bộ dữ liệu như Wine và VRB, thuật toán đề xuất cho tập rút gọn có kích thước nhỏ hơn nhưng khả năng phân loại chính xác lại cao hơn so với các thuật toán khác. Kết quả này càng khẳng định, phương pháp xác định độ quan trọng của thuộc tính theo tiếp cận tính toán biên trên cấu trúc tôpô đề xuất cho tập rút gọn hiệu quả về kích thước và khả năng phân loại chính xác.



Hình 1. Mối quan hệ của kích thước của tập rút gọn bởi mỗi thuật toán và khả năng phân loại chính xác trên SVM



Hình 2. Mối quan hệ của kích thước của tập rút gọn bởi mỗi thuật toán và khả năng phân loại chính xác trên KNN

VI. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã đề xuất phương pháp rút gọn thuộc tính theo tiếp cận tính toán biên của tập mờ. Trong đó, cấu trúc tập mờ được xây dựng dựa trên quan hệ thứ tự của các phép toán xấp xỉ mờ trên và xấp xỉ mờ dưới của tập thô mờ. Trên cơ sở đó chúng tôi đề xuất mô hình rút gọn thuộc tính theo tiếp cận tập mờ, trong đó phương pháp xác định độ quan trọng của thuộc tính dựa trên tiếp cận tính toán biên và tập rút gọn được định nghĩa theo tiếp cận bảo toàn của cấu trúc tập mờ đề xuất. Các kết quả thực nghiệm cho thấy thuật toán đề xuất cho tập rút gọn hiệu quả về kích thước và khả năng phân loại chính xác. Tuy nhiên thời gian thực hiện của thuật toán còn hạn chế. Kết quả của nghiên cứu này đóng góp một góc nhìn mới về tập rút gọn theo tiếp cận tập mờ trên cơ sở mối quan hệ của tập xấp xỉ mờ trên xấp xỉ mờ dưới của tập thô mờ.

Trong tương lai, chúng tôi tiếp tục phát triển mô hình rút gọn thuộc tính này trong các trường hợp bảng dữ liệu thay đổi về thuộc tính và đối tượng. Bên cạnh đó, các tính chất quan hệ của tập xấp xỉ trên và xấp xỉ dưới trong tập mờ cũng được nghiên cứu và khảo sát để giảm thời gian tính toán.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi Quỹ Phát triển khoa học và công nghệ Quốc gia (NAFOSTED) trong Đề tài mã số 102.05-2021.10.

TÀI LIỆU THAM KHẢO

[1] J. Ye, J. Zhan, W. Ding, and H. Fujita, "A novel fuzzy rough set model with fuzzy neighborhood operators," *Inf. Sci. (Ny)*, vol. 544, pp. 266-297, 2021, DOI: 10.1016/j.ins.2020.07.030.

- [2] S. An, Q. Hu, and C. Wang, "Probability granular distance-based fuzzy rough set model," *Appl. Soft Comput.*, vol. 102, 2021, DOI: 10.1016/j.asoc.2020.107064.
- [3] C. Wang, Y. Huang, M. Shao, and X. Fan, "Fuzzy rough set-based attribute reduction using distance measures," *Knowledge-Based Syst.*, vol. 164, pp. 205-212, 2019, DOI: 10.1016/j.knosys.2018.10.038.
- [4] T. T. Nguyen *et al.*, "A novel filter-wrapper algorithm on intuitionistic fuzzy set for attribute reduction from decision tables," *Int. J. Data Warehous. Min.*, vol. 17, no. 4, pp. 67-100, 2021, DOI: 10.4018/IJDWM.2021100104.
- [5] N. L. Giang *et al.*, "Novel Incremental Algorithms for Attribute Reduction from Dynamic Decision Tables Using Hybrid Filter-Wrapper with Fuzzy Partition Distance," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 5, pp. 858-873, 2020, DOI: 10.1109/TFUZZ.2019.2948586.
- [6] E. F. Lashin and T. Medhat, "Topological reduction of information systems," *Chaos, Solitons and Fractals*, vol. 25, no. 2, pp. 277-286, 2005, DOI: 10.1016/j.chaos.2004.09.107.
- [7] K. Qin and Z. Pei, "On the topological properties of fuzzy rough sets," *Fuzzy Sets Syst.*, vol. 151, no. 3, pp. 601-613, 2005, DOI: 10.1016/j.fss.2004.08.017.
- [8] H. Yu and W. R. Zhan, "On the topological properties of generalized rough sets," *Inf. Sci. (Ny)*, vol. 263, pp. 141-152, 2014, DOI: 10.1016/j.ins.2013.09.040.
- [9] Y. Xie and X. Gao, "Topological reduction algorithm for relation systems," *Soft Comput.*, vol. 26, no. 22, pp. 11961-11971, 2022, DOI: 10.1007/s00500-022-07431-y.
- [10] S. Mishra and R. Srivastava, "Fuzzy topologies generated by fuzzy relations," *Soft Comput.*, vol. 22, no. 2, pp. 373-385, 2018, DOI: 10.1007/s00500-016-2458-6.
- [11] Z. Li, T. Xie, and Q. Li, "Topological structure of generalized rough sets," *Comput. Math. with Appl.*, vol. 63, no. 6, pp. 1066-1071, 2012, DOI: 10.1016/j.camwa.2011.12.011.
- [12] H. Wu and G. Liu, "The relationships between topologies and generalized rough sets," *Int. J. Approx. Reason.*, vol. 119, pp. 313-324, 2020, DOI: 10.1016/j.ijar.2020.01.011.
- [13] L. Su and W. Zhu, "Dependence space of topology and its application to attribute reduction," *Int. J. Mach. Learn. Cybern.*, vol. 9, no. 4, pp. 691-698, 2018, DOI: 10.1007/s13042-016-0598-8.
- [14] Z. Pei, D. Pei, and L. Zheng, "Topology vs generalized rough sets," *Int. J. Approx. Reason.*, vol. 52, no. 2, pp. 231-239, 2011, DOI: 10.1016/j.ijar.2010.07.010.
- [15] L. Yang and L. Xu, "Topological properties of generalized approximation spaces," *Inf. Sci. (Ny)*, vol. 181, no. 17, 2011, DOI: 10.1016/j.ins.2011.04.020.
- [16] UCI. machine learning repository (2021). Data. <https://archive.ics.uci.edu/ml/index.php>.

ROUGH SET BASED TOPOLOGY STRUCTURE FOR ATTRIBUTE REDUCTION IN DECISION TABLE

Tran Thanh Dai, Nguyen Long Giang, Vu Duc Thi, Tran Thi Hue

ABSTRACT: Attribute reduction is an essential data preprocessing step in machine learning and data mining. In recent years, the relationship between the rough set theory model and topology and methods of building topological structures according to the rough set approach has received much attention from researchers. However, building topological structures according to a rough set approach for attribute reduction problems still faces many challenges. In this study, we propose to build a topological structure based on the ordinal relationship of the upper and lower approximations of the fuzzy rough set. We propose an attribute reduction method according to the boundary calculation approach of the proposed topology applied to the attribute reduction problem.